

Support Vector Clustering Preliminary Experiments

Università degli studi di Napoli “Federico II”

Via Cinthia, 80126, Napoli, Italy

Vincenzo Russo
(vincenzo.russo@neminis.org)

July 4, 2007

Abstract

In this document we present the results of some preliminary experiments using Support Vector Clustering (SVC). The experiments were conducted both over synthetic data sets and over real-world data sets taken from UCI Repository of Machine Learning Database.

Contents

1	Hardware	3
2	Software	3
3	Clustering evaluation	3
4	Real-world 1: The Iris Plant Database	4
4.1	The Data Set	4
4.2	The results	4
4.2.1	Conclusion	5
5	Synthetic 1: Syndeca 02	5
5.1	The Data Set	5
5.2	The results	5
5.2.1	Conclusion	6

1 Hardware

The hardware configuration used for tests:

- CPU: PowerPC G4 at 1.5GHz
- RAM: 768 MBytes
- OS: Mac OS X

2 Software

Since SVC is a young approach to clustering [BHHSV01, BHSHV00, BHHSV00], there is no software available, especially for the second stage of the algorithm, namely *Cluster Labeling*. So, we started to develop an Object Oriented implementation of the Support Vector Clustering, in order to implement in an elegant way any variation of the original proposal. The code is written in C++ and relies on LIBSVM library [CL01] for the first stage of the SVC, namely the *Domain Description* [SPST⁺99, TD99a, TD99b, TAX01]. All the remaining stuff was developed *ex novo*. We have implemented two types of cluster labeling algorithms, the original one [BHHSV01] (also known as *Complete Graph Cluster Labeling*) and the more promising one in literature, namely the *Cone Cluster Labeling* [LD06, LD05b], which is the faster one available and very accurate at the same time. Furthermore, a kernel width selection method [LD05a, LD05b] for the gaussian kernel is under development, which is close to the end of the first working release.

The tests were conducted using only the more efficient *Cone Cluster Labeling*, due hardware limitations.

3 Clustering evaluation

Due the nature of these preliminary tests, where for each data-set used we know also the original classification, we adopt the classic instruments used for evaluating classifications:

- Precision/Recall/ F_1 in the case of binary classification
- Macroaveraging in the case of multi-class

Furthermore, we also use the *accuracy* metric, although sometimes it is not an appropriate measure [MRS07, par. 8.3]; this is why it is often used in several papers evaluating their own clustering techniques.

4 Real-world 1: The Iris Plant Database

The *Iris Data Set* [Fis36] is the best known database to be found in the pattern recognition literature. The data set contains three classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other two; the latter are *not* linearly separable from each other.

4.1 The Data Set

The data-set was taken from *UCI Repository of Machine Learning Database* [NHBM98]:

- Number of instances: 150
- Number of attributes: 4
- Number of classes: 3
- Classes: Iris Setosa, Iris Versicolour, Iris Virginica
- Class distribution: 50 items per class (the first 50 items in the first class, and so on...)
- Peculiarity: 1 class is linearly separable from the other 2, but the other 2 are not linearly separable from each other

4.2 The results

We obtain a **macro-averaging** value of 89.0531% from the results in Table 1. The **accuracy** is 89.3333%.

Class (<i>vs Rest</i>)	Precision	Recall	F_1
1	100%	100%	100%
2	75.7576%	100%	86.2069%
3	100%	68%	80.9524%

Table 1: The precision/recall/f1 values obtained for each class of the Iris data-set, using SVC with Cone Cluster Labeling

4.2.1 Conclusion

We obtain quite good results, which are similar or better of the other classic algorithms results. Compared with results in several other papers about SVC [LD05b, BHHSV01], we note that our results compares worse with them, but a more deep analysis will disclose that in those papers the best results were obtained reducing the dimensionality of data-set (from 4D to 2D) through PCA. In fact, in [BHHSV01] the authors said that with complete data-set 14 misclassifications occur. In our case similar results are obtained, with 16 misclassifications. The difference could lie on the different Cluster Labeling algorithm and/or on the different implementation of the Domain Description part of the SVC. Furthermore, we recall our software is in an alpha status and it needs several improvements and corrections.

5 Synthetic 1: Syndeca 02

The first experiment on synthetic data was performed on a data-set created with *SynDECA* [VV05].

5.1 The Data Set

- Number of instances: 1000
- Number of attributes: 10
- Number of classes: 5
- Class distribution: 327 in the first class, 134 in the second one, 162 in the third one, 132 in the fourth one, 133 in the fifth one.
- Peculiarity: only 888 points are distributed over the five classes. Remaining 112 points are noise (outliers).
- Peculiarity: the various clusters have different shapes, like circle, rectangular, ellipse, random and square.

5.2 The results

We obtain a **macro-averaging** value of 100% and the **accuracy** is 100%.

5.2.1 Conclusion

The SVC has classified correctly all classifiable points. In addition, in this test we can observe one of the unique peculiarity of the SVC: the outliers handling. In fact, as mentioned above, this data-set has 112 outliers. The SVC detected all of them and clustered in a singleton cluster each. Next, it takes in account only non-singleton clusters, resulting in 100% of accuracy and macro-averaging. In this test SVC also shows its ability of dealing with any-shaped clusters.

References

- [BHHSV00] Asa Ben-Hur, David Horn, Hava T. Siegelmann, and Vladimir Vapnik. A support vector method for clustering. In *Neural Information Processing Systems*, pages 367–373, 2000. 2
- [BHHSV01] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik. Support vector clustering. *Journal of Machine Learning Research*, 2:125–137, 2001. 2, 4.2.1
- [BHSHV00] Asa Ben-Hur, Hava T. Siegelmann, David Horn, and Vladimir Vapnik. A support vector clustering method. *International Conference on Pattern Recognition*, 02:2724, 2000. 2
- [CL01] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines, 2001. Manual available at <http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf>. 2
- [Fis36] Ronald Aylmer Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936. 4
- [LD05a] Sei-Hyung Lee and Karen M. Daniels. Gaussian kernel width generator for support vector clustering. In Matthew He, Giri Narasimhan, and Sergei Petoukhov, editors, *Advances in Bioinformatics and Its Applications*, volume 8, pages 151–162, 2005. 2
- [LD05b] Sei-Hyung Lee and Karen M. Daniels. Gaussian kernel width selection and fast cluster labeling for support vector clustering. Technical report, Department of Computer Science, University of Massachusetts Lowell, 2005. 2, 4.2.1
- [LD06] Sei-Hyung Lee and Karen M. Daniels. Cone cluster labeling for support vector clustering. In *Proceedings of 6th SIAM Conference on Data Mining*, pages 484–488, May 2006. 2
- [MRS07] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2007. 3
- [NHBM98] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. UCI repository of machine learning databases, 1998. 4.1

- [SPST⁺99] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. Technical Report 99-87, Microsoft Research, Redmond, WA, 1999. 2
- [TAX01] David Martinus Johannes TAX. *One-class classification: concept learning in the absence of counter-examples*. PhD thesis, Technische Universiteit Delft, 2001. 2
- [TD99a] David M. J. Tax and Robert P. W. Duin. Data domain description using support vectors. In *European Symposium on Artificial Neural Network*, pages 251–256, Bruges (Belgium), April 1999. 2
- [TD99b] David M. J. Tax and Robert P. W. Duin. Support vector domain description. *Pattern Recognition Letters*, 20(11-13):1191–1199, 1999. 2
- [VV05] Jhansi Rani Vennam and Soujanya Vadapalli. Syndeca: A tool to generate synthetic datasets for evaluation of clustering algorithms. In *11th International Conference on Management of Data (COMAD 2005)*, Goa, India, January 2005. <http://cde.iiit.ac.in/syndeca>. 5